Effectiveness and Efficiency for Document Clustering in Biomedicine

Kazuhiro Seki
Faculty of Intelligence and Informatics
Konan University
Kobe, Japan
seki@konan-u.ac.jp

Michael Segundo Ortiz
Carolina Health Informatics Program
University of North Carolina
Chapel Hill, USA
msortiz@unc.edu

Javed Mostafa
School of Information and Library Science
University of North Carolina
Chapel Hill, USA
jm@unc.edu

***Abstract*—It is crucial for biomedical information retrieval, or clinical decision support in particular, to discover relevant biomedical/clinical information buried in scientific publications. At present, typical search interface is based on keywords as queries and returns a ranked list of documents, which is suited for finding simple factoids but not ideal for more complex information needs required for clinical decision support. A search interface deemed more suitable for this kind of tasks is cluster-based browsing, where retrieved documents are topically grouped for more intuitive exploration. To adopt this model, however, one needs to consider not only the effectiveness but also the efficiency of clustering framework as clustering is a computationally costly operation. As a first step toward a cluster-based browsing information exploration, this paper empirically studies representative feature selection/extraction methods and clustering algorithms for their effectiveness and efficiency.**

***Index Terms*—information retrieval, document representation, clustering, effectiveness, efficiency**

I. Introduction

PubMed relies on a keyword-based query and ranked-retrieval model, which is unarguably a useful tool for finding information but also poses a limitation as to effective information exploration. As an example of the ineffectiveness of conventional search, consider that a single search on PubMed for “breast cancer” returns a long list of nearly 400,000 articles as of October 2019, ranging from basic research on cell culture and genetic polymorphisms to others covering treatments and pharmacological responses. If it were a factoid search asking, one could find the right information among the top ranked results. However, researchers’ or clinicians’ interest is often broader and requires them to go through the ranked list to find all the relevant information.

A possible remedy to this situation is to employ an information retrieval (IR) system with a dynamic cluster-based browsing interface. Such an IR system takes either the entire literature or a search result for a particular query and identifies topical clusters of documents. These systems can be equipped with functionalities for interaction and visualization, which potentially allow us to effectively find relevant documents by visually browsing and iteratively navigating a vast document collection.

Seminal work on the cluster-based IR paradigm, called Scatter/Gather, was conducted by Cutting et al. [1] as an alternative to the standard keyword-based search. Recently, Ortiz et al. [2] adopted the Scatter/Gather model and developed a prototype system to explore the biomedical literature and presented a use case describing how the system could help the user navigate a document collection. However, the particular work is more of system development and does not discuss the quality of the clusters.

In the past, some studies investigated cluster quality and the relationship to IR [3]–[5]. However, they focused more on algorithmic complexity, visual interfaces and cognitive load, cluster relevance feedback based on user profiles, and user modeling to guide adaptive visualizations. On the other hand, our work is specifically targeting the biomedical domain and focuses on identifying the best document representation and clustering in terms of effectiveness and efficiency. Specifically, we empirically compare representative feature selection/extraction methods and clustering algorithms on a standard benchmark data set and a newly created data set derived from Text Retrieval Conference (TREC) Precision Medicine (PM) track topics.[1]

II. Methodology

A. Data sets

Our goal of this work is to identify the most effective and efficient automated clustering approach for biomedical documents. To evaluate the effectiveness of clustering approach, cluster quality needs to be evaluated by some external criteria [6], which however requires ground truth cluster labels. For this purpose, we take advantage of subject headings (i.e., MeSH terms) as a proxy for cluster labels.

We used two sets of data sets: one for feature selection/extraction and the other for clustering. The former includes a subset of OHSUMED [7], which was compiled from PubMed as a benchmark data set originally for information filtering task. Joachim [8] later used a part of the data for

Supported by the NIH-NLM T15 grant 5T15LM012500-02, United Health Foundation’s ENABLE grant, JSPS KAKENHI Grant Number JP18K11558, and MEXT, Japan.

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

[1] http://www.trec-cds.org/2018.html

evaluating supervised classifiers using 23 MeSH terms under the MeSH "disease" category as class labels, of which we used the four most frequent MeSH terms ("Neoplasms", "Nervous System Diseases", "Cardiovascular Diseases", and "Pathological Conditions, Signs and Symptoms") to facilitate evaluation.

As for clustering experiments, we compiled a new, larger data set based on the 50 topics provided for the TREC PM track 2018. Among the 50 topics, there were 22 unique disease names and we retrieved a set of articles with each disease name being a query on August 8th and 9th in 2019 on the PubMed search interface.

For the set of resulting articles for each disease, we identified the 10 most frequent major MeSH terms and retained only the articles which have any of the 10 MeSH terms for simplicity. We expect that the restriction to major MeSH terms yields a document collection more focused on certain topics, which would be more appropriate as underlying topics (i.e., cluster labels). In addition, among the resulting 22 articles sets for 22 diseases, we use only 10 smaller sets as some of the clustering algorithms we examined did not scale to larger data sets. The resulting data set is hereafter called the "10 disease data set".

B. Document representation

For document representation, we use only titles and abstracts of articles and apply the bag-of-words vector space model to construct a term-document matrix M with tfidf term weighting [6].

For feature selection/extraction, we compare two alternative methods, namely, document frequency (DF) thresholding [9] and singular value decomposition (SVD) [11]. The former is a feature selection method and set a predefined threshold, denoted as τ_{df}, to pick only terms whose df is greater than the threshold, whereas SVD [11] is a feature extraction method to generate new features based on matrix decomposition.

After applying one of the feature selection/extraction methods, the resulting term-document matrix M is fed to a clustering algorithm. It should be emphasized that the MeSH terms are only used for evaluation purpose and not included in a term-document matrix.

C. Clustering algorithms

For clustering, we examine the following five algorithms: mini-batch k-means [12], spectral clustering [13], non-negative matrix factorization (NMF) [14], and SVD (also known as latent semantic analysis; LSA). The last two algorithms are not particularly for clustering but can be seen as soft clustering. We briefly describe each algorithm in the following except for SVD already introduced in the previous section.

Mini-batch k-means uses a mini-batch optimization for k-means clustering, which greatly reduces computation time but still achieves a solution close to the standard k-means algorithm. The algorithm first takes b random samples as a mini-batch and each sample in the mini-batch is assigned with the nearest centroid and then each centroid is updated per-sample basis. The assignment and update steps are repeated for predetermined times or until convergence. In the following, we use "k-means" to refer to mini-batch k-means for short.

Spectral clustering also uses standard k-means clustering but not on the term-document matrix M' but on k largest eigenvectors of a Laplacian matrix of A, where A is a similarity matrix of M'. By design, this clustering algorithm is effective for cases where natural clusters of given data are not in convex regions.

NMF is a matrix factorization algorithm to find non-negative factors W and H for a given non-negative matrix V, i.e., $V \approx WH$, subject to $W \geq 0$, $H \geq 0$. Let V be an $l \times m$ term-document matrix with l and m being the number of terms and the number of documents, respectively, and k be the number of clusters. Applying NMF to V yields $l \times k$ matrix W and $k \times m$ matrix H. Then, the latter matrix gives (soft) cluster membership and we choose $\arg\max_i H_{ij}$ as the cluster label for document d_j.

III. Evaluation

A. Experimental setup

There are a few parameters involved in our clustering framework. We examined the following values for the parameters: DF threshold $\tau_{df} = \{1, 2, 4, 7, 10, 20, 50, 100, 200\}$ and the number of dimensions (components) for SVD $n = \{6, 8, 10, 20, 50, 100, 200, 400, 700, 1000\}$. The number of clusters was fixed to four for the OHSUMED data set and to 10 for the 10 disease data set, considering the underlying topic sets.

For evaluating cluster quality, we used adjusted mutual information (AMI) as the primary metric. In addition, we reported purity quantifying how homogeneous each cluster is. Note that perfect purity (= 1) can be easily achieved when each document forms its own cluster.

All experiments for feature selection/extraction (Section III-B) were carried out on a computer with dual-core 1.6 GHz Intel Core i5 processor using 16 GB RAM, and for clustering (Section III-C) on a computer with two 6-core 2.10 GHz Intel Xeon E5-2620 v4 processors using 192 GB RAM.

B. Feature selection/extraction

We compared the two feature selection/extraction methods with k-means. Fig. 1 and Fig. 2 summarize the results for different parameter values.

Fig. 1 indicates that DF thresholding is not effective when the threshold $\tau_{df} > 10$. In terms of efficiency, the total computation time was found smaller in the beginning, then increased gradually up to $\tau_{df} \approx 50$. Taken together, it is recommended that τ_{df} should be small, around 2 to 10.

Then, Fig. 2 indicates that clustering performance was relatively high when the number of dimensions n was low (around 10 to 20), and it rapidly decreased as n increased up to 200, and then it again went up with greater n. Note that while further increasing n may be beneficial for clustering performance, SVD with large dimensions is time-consuming

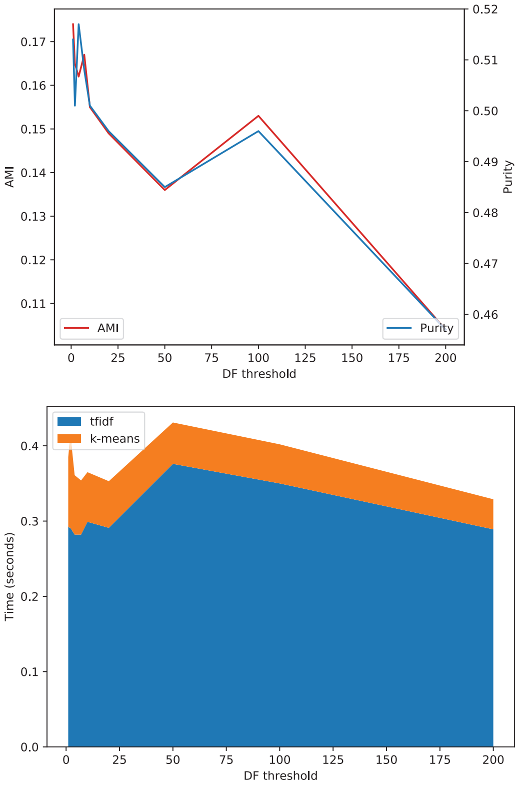

Fig. 1. Relation between DF threshold τ_{df} and clustering performance (top) and computation time (bottom) on the OHSUMED data.

TABLE I
CLUSTER PERFORMANCE AND COMPUTATION TIME FOR DIFFERENT CLUSTERING ALGORITHMS ON THE 10 DISEASES DATA SET.

Clustering	Time	AMI	Purity
SVD	0.2304 (0.0045)	0.1140 (0.0016)	0.5708 (0.0009)
k-means	0.4207 (0.0385)	0.1891 (0.0144)	0.6224 (0.0134)
NMF	2.6633 (0.0683)	0.1850 (0.0004)	0.6229 (0.0007)
spectral	20.8130 (0.3886)	0.1750 (0.0010)	0.6091 (0.0008)

and takes up most of the total processing time. All in all, SVD does not seem to offer much advantage as a feature extraction method on the OHSUMED data.

C. Clustering

Next, we applied four clustering algorithms described in Section II-C to the 10 diseases data set. Table I compared them in terms of clustering performance in AMI and purity and computation time averaged over 10 diseases and over 20 trials for each disease, where rows were sorted in ascending order of computation time. The figures in the parentheses are the averages of standard deviation. The computation time was only for clustering, excluding preprocessing such as selection, where τ_{df} was set to 2.

The result shows that k-means and NMF are the best algorithms in terms of clustering performance, although the latter is six times slower than k-means. For computation time, SVD was the fastest while its performance was inferior.

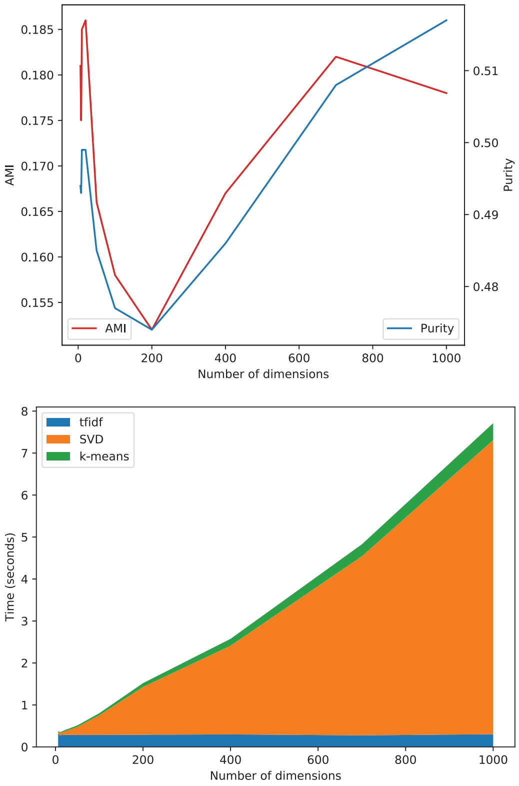

Fig. 2. Relation between the number of dimensions n for SVD and clustering performance (top) and computation time (bottom) on the OHSUMED data.

Overall, k-means appears to be the best choice balancing effectiveness and efficiency.

IV. CONCLUSIONS

As a first step toward cluster-based browsing for the biomedical literature, we empirically investigated the effectiveness and efficiency of feature selection/extraction methods and clustering algorithms. Specifically, we compared DF thresholding and SVD for feature selection/extraction on a subset of the OHSUMED data and found that the performance improvement for clustering was marginal.

For comparing clustering algorithms, we constructed a data set focusing on 10 diseases from TREC PM topics, where frequent major MeSH terms were used as ground truth. Our experiments showed that mini-batch k-means and NMF were among the best in terms of cluster quality. In term of efficiency, SVD was found the fastest, followed by mini-batch k-means.

Building upon this work, we are currently developing a dynamic, cluster-based information exploration. A prototype system can be accessed at our project website.[2]

REFERENCES

[1] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections," in *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1992, pp. 318–329.

[2] http://pattie.unc.edu

[2] M. Ortiz, H. Kim, M. Wang, K. Seki, and J. Mostafa, "Dynamic cluster-based retrieval and discovery for biomedical literature," in *Proceedings of the 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB)*, 2019.

[3] J.-W. Ahn and P. Brusilovsky, "Adaptive visualization for exploratory information retrieval," *Information Processing & Management*, vol. 49, no. 5, pp. 1139–1164, 2013.

[4] W. Ke, C. R. Sugimoto, and J. Mostafa, "Dynamicity vs. effectiveness: studying online clustering for scatter/gather," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 19–26.

[5] J. Zhang, J. Mostafa, and H. Tripathy, "Information retrieval by semantic analysis and visualization of the concept space of d-lib® magazine," *D-lib Magazine*, vol. 8, no. 10, pp. 1082–9873, 2002.

[6] C. D. Manning, P. Raghavan, and H. Schütze, "others, introduction to information retrieval, vol. 1," 2008.

[7] W. Hersh, C. Buckley, T. J. Leone, and D. Hickam, "OHSUMED: An interactive retrieval evaluation and new large test collection for research," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '94. New York, NY, USA: Springer-Verlag New York, Inc., 1994, pp. 192–201. [Online]. Available: http://dl.acm.org/citation.cfm?id=188490.188557

[8] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, ser. ECML '98. London, UK, UK: Springer-Verlag, 1998, pp. 137–142. [Online]. Available: http://dl.acm.org/citation.cfm?id=645326.649721

[9] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning*, 1997, pp. 412–420.

[10] J. Mostafa, L. M. Quiroga, and M. Palakal, "Filtering medical documents using automated and human classification methods," *Journal of the American Society for Information Science*, vol. 49, no. 14, pp. 1304–1318, 1998.

[11] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[12] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1177–1178. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772862

[13] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, ser. NIPS'01. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856. [Online]. Available: http://dl.acm.org/citation.cfm?id=2980539.2980649

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, ser. NIPS'00. Cambridge, MA, USA: MIT Press, 2000, pp. 535–541. [Online]. Available: http://dl.acm.org/citation.cfm?id=3008751.3008829